REMARKS

Claims 1, 19, 20, 21, 28, 29, and 34 have been amended. Claims 35 – 37 have been deleted. Hence, Claims 1 - 34 are pending in the Application.

As a preliminary matter, receipt of the Notice of Draftsperson's Patent Drawing Review is acknowledged. The Office Action has stated that the informal drawings submitted are suitable only for prosecution, and that formal drawings must be submitted upon allowance. Therefore, Applicant will address the issues identified in the Notice of Draftsperson's Patent Drawing Review when formal drawings are submitted after allowance.

SUMMARY OF REJECTIONS/OBJECTIONS

Claims 19 – 25 are rejected under 35 USC 112, first paragraph, as containing subject matter not described in the specification in an enabling way because the specification does not describe what in the matrix constitutes structural support.

Claims 19 – 25 are rejected under USC 112, second paragraph, as being indefinite for failing to point at and distinctly claim the subject matter which applicant regards as the invention because the term structural information lacks antecedent basis.

Claims 28 – 32 are rejected under USC 112, first paragraph, as containing subject matter not described in the specification in an enabling way because the specification does not describe the operation or construction of a second matrix.

Claims 28 – 32 are rejected under USC 112, second paragraph, as being indefinite for failing to point out and distinctly claim the subject matter which applicant regards as the invention because the term "second matrix" lacks antecedent basis.

Claims 1 – 12, 19, 26, and 33 –37 are rejected under USC 102(a) and (e) as being anticipated by U.S. Patent Number 5,835,905, herein after *Pirolli*.

Claims 13 – 14, 17 – 18, and 27 – 32 are rejected under 35 USC 103(a) as being unpatentable over *Pirolli*.

Claim 20 is rejected under USC 103(a) as being unpatentable over *Pirolli* in view of U.S. Patent No. 6,128,606, hereinafter *Bengio*.

Claims 21 – 25 are rejected under USC 103(a) as being unpatentable over *Pirolli* in view of U.S. Patent No. 6,389,436, hereinafter *Chakrabarti*.

REJECTIONS UNDER USC 112

Claims 19 – 25 and 28 – 32, as amended, provide antecedent basis for the elements requiring it. Support for the amended limitations may be found at page 22, line 7. Removal of the rejections under USC 112 is respectfully requested.

REJECTION OF CLAIMS 1 AND 34 UNDER USC 102

Claims 1 and 34 recite:

establishing a plurality of training sets, wherein each training set is associated with a category and includes training documents that have been classified as belonging to said associated category;

determining how strongly each document of said plurality of documents corresponds to each of said plurality of categories by determining similarity between said each document and the training documents that belong to the training set of said category; and

wherein the step of determining similarity is performed using a matrix representing document similarity that is derived by combining two or more measures of document similarity.

Claims 1 and 34 recite limitations not disclosed or suggested by the cited art. Among these limitations are "determining how strongly each document of said plurality

of documents corresponds to each of said plurality of categories by determining similarity between said each document and the documents that belong to the training set of said category."

The Office Action has based the rejection of claims 1 and 34 on *Pirolli*. While *Pirolli* discloses categorizing a set of documents according to "classification characteristics", it fails to disclose the particular way of categorizing claimed, that is, it fails to disclose classifying documents into a category based on similarity to a set of pre-classified training documents that have been classified into that category. The category for which the strength of correspondence to a document is being determined is referred to in this response as the target category.

Pirolli teaches that "The classification characteristics are predetermined rules which are applied to the feature vectors of the page.... When the classification characteristics are applied to the respective feature vectors, lists of pages in the particular classes are created." (col. 5, lines 12 – 23) "In order to perform categorizations each Web page at the Web locality is represented by a vector of features constructed from the above topology, meta-information, usage statistics and paths, and text similarities. These Web page vectors are collected into a matrix. Such a matrix is illustrated in FIG. 5. Referring to FIG. 5, each row 501 of the matrix 500 represents a Web page. The columns in matrix 500 represent a the page's:

page identifier, identifies the particular web page (column 502).

size, in bytes, of the item (column 503) inlinks, the number of hyperlinks that point to the item from the web locality (column 504).

outlinks, the number of hyperlinks the item contains that point to other items in the web locality (column 505).

frequency, the number of times the item was requested in the sample period (column 506).

sources, number of times the item was identified as the start of a path traversal (column 507).

csim, the textual similarity of the item to it's children based upon previous SCA calculation (column 508).

cdepth, the average depth of the item's children as measured by the number of '/' in the URL (column 509)." (col. 8, lines 7 – 32)

To disclose the limitation in issue, *Pirolli* must not only teach that a set of documents are classified within a category based on some features, but also that:

(1) a determination is being made about how strongly a set of document corresponds to a target category, where the determination is based on similarity with another set of documents; and

(2) the other set of documents constitute a training set comprised of training documents that have already been classified as belonging to the target category.

As shown above, *Pirolli* teaches that the classification of documents is based on a variety of features. These features, with the exception of one, are not based in any way upon document similarity. Therefore, these features (minus the exception) can not disclose or in anyway suggest a determination being made about how strongly documents correspond to a category based on similarity with another set of documents, as claimed.

The one feature that *Pirolli* teaches is based on document similarity is the csim feature, which measures similarity between the document to be classified and its children. One could allege that the children constitute a set comprised of training documents that have been classified into the category of children of the document being classified.

However, with respect to the training documents in the training set, claims 1 and 34 do not merely require that the training documents belong to any given category, but that they belong to the target category, the category whose strength of correspondence with a plurality of documents is being determined by the method claimed.

To suggest the training set as claimed, *Pirolli* could teach to determine how strongly a document belongs to the category of its children by determining the document's similarity to its children. *Pirolli*, however, does not teach or suggest such a step.

As shown above, *Pirolli* does not disclose or suggest all the features of claims 1 and 34, and does not anticipate or render obvious all limitations of claims 1 and 34. Therefore, claims 1 and 34 are patentable. Reconsideration and allowance of claims 1 and 34 is respectfully requested.

PENDING CLAIMS

The pending claims not discussed so far are dependant claims that depend on an independent claim that is discussed above. Because each of the dependant claims include the limitations of claims upon which they depend, the dependant claims are patentable for at least those reasons the claims upon which the dependant claims depend are patentable. Removal of the rejections with respect to the dependant claims and allowance of the dependant claims is respectfully requested. In addition, the dependent claims introduce additional limitations that independently render them patentable. Due to the fundamental difference already identified, a separate discussion of those limitations is not included at this time.

The Examiner is respectfully requested to contact the undersigned by telephone if it is believed that such contact would further the examination of the present application.

For the reasons set forth above, Applicant respectfully submits that all pending claims are patentable over the art of record, including the art cited but not applied. Accordingly, allowance of all claims is hereby respectfully solicited.

Respectfully submitted,

HICKMAN PALERMO TRUONG & BECKER LLP

Dated: January 9, 2003

Marcel K. Bingham
Reg. No. 42,327

1600 Willow Street
San Jose, CA 95125
Telephone No.: (408) 414-1080 ext.206
Facsimile No.: (408) 414-1076

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner for Patents, Washington, D.C. 20231

on January 9, 2003 by Trudy Bigelow

Marked-Up Version of Amendments

In the Specification:

Replace the two paragraphs beginning at page 2 line 19 with the following:

New approaches exploiting hyperlink structure of the Web are addressing this problem. For example, the CLEVER project of the IBM Almaden Research Center, San Jose, California is developing a search engine and directory engine that is described in documents available at [~~http://www.almaden.ibm.com/cs/k53/clever.htm~~] "http://www.almaden.ibm.com/cs/k53/clever.htm". This work is based on the Hypertext Induced Topic Search process developed by Jon Kleinberg. Generally, in this process, a standard text search engine generates a Training Set of electronic documents that match a query subject or category. The process extends the Training Set to include all documents pointing to or pointed to by each document in the Training Set. Using information that describes the links between the documents, the process seeks the best Authorities and Hubs that match the query or category. Mathematically, the Authorities and Hubs are the principal Eigenvectors of matrices representing the link relations between the documents.

In another approach, the GOOGLE project [~~(http://google.stanford.edu)~~] (at URL google.stanford.edu) uses a process of generating PageRanks. PageRanks are iteratively updated based on linked hypertext structures. The resulting PageRanks measure the general connectedness of documents on the Web, without regard to a particular category or query. The assumption is that more connected documents will tend to be of general interest to most users.

Replace the paragraph on page 12, line 17 with:

For instance, if one document is clicked often in the same time period as another, and both documents are viewed for a long period of time, the classification system may determine that the two pages are relevant to the user's interests and have Similarity. Two

assumptions are made. First, the frequency at which a user changes his or her topic of interest is much slower than [~~that~~] the frequency at which a user changes pages. Second, a user's interest in a page can be estimated by a function of the time that the user spends in viewing the page.

Replace the paragraph on page 14, line 3 with:

[~~6.~~]5. TITLE SIMILARITY

Replace the paragraph on page 14, line 16 with:

[~~7.~~]6. URL SIMILARITY

Replace the paragraph on page 15, line 4 with:

[~~8.~~]7. CACHE HIT LOG SIMILARITY

Replace the paragraph beginning at page 15, line 20 with:

A human user of the client 200, or an agent executing in the client, instructs browser 202 to request a hypertext document according to a particular location identifier. For example, a Web browser of the client may request a Web document using its URL, such as [~~"http://www.inktomi.com/"~~]-"www.inktomi.com/". Browser 202 submits the request to cache server 208. The cache server 208 determines whether the requested document is already in the cache 210. If it is, the cache server 208 delivers the requested document to the browser 202 from the cache 210. If it is not, cache server 208 uses a domain name service or similar network element of network 204 to determine the location of origin server 220. Cache server 208 then requests the document from origin server 220, via network 204. Finally, cache server 208 stores a copy of the document in cache 210, and passes a copy of the document back to browser 202. Thus, normally, all Web traffic directed from browser 202 passes through the cache server 208. The cache server is thereby in the ideal position to log users' requests for various Web documents.

Replace the paragraphs beginning at page 16, line 23:

14-Feb-1999 08:01:22, 255.1.2.254, [~~http://www.inktomi.com/index.html~~] "www.inktomi.com/index.html"

14-Feb-1999 08:01:23, 199.22.131.44, [~~http://www.mwe.com/~~] "www.mwe.com/"

14-Feb-1999 08:01:26, 255.1.2.254, [~~http://www.inktomi.com/products~~] "www.inktomi.com/products"

14-Feb-1999 08:01:27, 199.22.131.44, [~~http://www.mwe.com/bios~~] "www.mwe.com/bios"

14-Feb-1999 08:01:31, 255.1.2.254, [~~http://www.inktomi.com/products/trafficserver.html~~] "www.inktomi.com/products/trafficserver.html"

14-Feb-1999 08:01:32, 199.22.131.44, [~~http://www.mwe/com/bios/palec.htm~~] "www.mwe/com/bios/palec.html"

Replace the abstract at page 40 line 2 with:

~~A method and apparatus are provided for determining when electronic documents stored in a large collection of documents are similar to one another. A plurality of similarity information is derived from the documents. The similarity information may be based on hyperlinks in the documents, text similarity, similarity of multimedia components in the documents, user click-through information, similarity in the titles of the documents or their location identifiers, etc. Another source of similarity information is patterns of user viewing, as may be monitored by a Web caching system. A pair of documents may be inferred to be similar if a particular user has shown high interest in both of them within a particular session or time period; alternatively, a pair of documents may be inferred to be similar if the pattern interest in the by all users for all sessions is similar. User interest is considered to be a function of the time that a user has viewed the document. The similarity information is fed to a combination function that synthesizes the various measures of similarity information into combined similarity information. Using the combined similarity information, an objective function is iteratively maximized in order to yield a generalized similarity value that expresses the similarity of particular pairs of documents. In an embodiment, the generalized similarity value is used to~~

~~determine the proper category, among a taxonomy of categories in an index, cache or search system, into which certain documents belong.~~

A method and apparatus are provided for determining when electronic documents stored in a large collection of documents are similar to one another. A plurality of similarity information is derived from the documents. The similarity information may be based on a variety of factors, including hyperlinks in the documents, text similarity, user click-through information, similarity in the titles of the documents or their location identifiers, and patterns of user viewing. The similarity information is fed to a combination function that synthesizes the various measures of similarity information into combined similarity information. Using the combined similarity information, an objective function is iteratively maximized in order to yield a generalized similarity value that expresses the similarity of particular pairs of documents. In an embodiment, the generalized similarity value is used to determine the proper category, among a taxonomy of categories in an index, cache or search system, into which certain documents belong.

In the Claims:

1. (Amended) A method of categorizing a plurality of new electronic documents into a set of categories, ~~each of~~[5] **comprising**[6] the ~~categories containing~~[7] **steps of:**[8]

establishing[9] a plurality of training **sets, wherein each training** [10]set **is associated with a category and includes training** [11]documents~~, by~~[12] **that have been classified as belonging to said associated category;**[13]

determining how strongly each document of said plurality of documents corresponds to each of said plurality of categories by determining similarity between said each document and the training documents that belong to the training set of said category; and[14]

wherein the step of determining similarity is performed[15] using a matrix representing document similarity that is derived by combining two or more measures of document similarity.

2. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity include hyperlink similarity.

3. (Unchanged) A method as recited in Claim 2, in which two documents among the plurality of documents are considered similar to each other when there is a link from one to the other, or when the two documents link to, or are linked to by, a set of other associated documents.

4. (Unchanged) A method as recited in Claim 3, in which certain hyperlinks have greater or lesser similarity weight than other hyperlinks, based on other features of the links or their source or destination documents.

5. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity include a similarity of text of the documents.

6. (Unchanged) A method as recited in Claim 5, wherein two documents are considered similar based on a comparison of word vectors derived from the text of each of the two documents.

7. (Unchanged) A method as recited in Claim 5, wherein text similarity is determined in part based upon weight values assigned to words of the text, and wherein certain words have greater or lesser weight than other words.

8. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity include user click-through similarity.

9. (Unchanged) A method as recited in Claim 8, wherein two documents are considered similar based on user click-through similarity when the documents are associated with similar patterns of user click behavior, selected from among frequency of clicks,

click context, duration of viewing, proximity in time to other clicks, or proximity in context to other clicks.

10. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity are derived from patterns detected in user viewing of the documents.

11. (Unchanged) A method as recited in Claim 10, wherein the user viewing information is monitored by a web caching system and stored in a log.

12. (Unchanged) A method as recited in Claim 10, wherein two documents are considered similar based on patterns of user viewing behavior, including frequency of viewing, viewing context, duration of viewing, proximity in time to other documents viewed by the same user, or similarity of patterns of viewing by all users.

13. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity include URL similarity.

14. (Unchanged) A method as recited in Claim 13, wherein two documents are considered similar if a URL of each document contains similar URL sub-components.

15. (Unchanged) A method as recited in Claim 1, wherein the measures of document similarity include multimedia similarity.

16. (Unchanged) A method as recited in Claim 15, wherein two documents are considered similar based on features derived from multimedia components linked to or contained by the documents.

17. (Unchanged) A method as recited in Claim 1, wherein the combination of two or more measures of document similarity is achieved by taking the union of each of a plurality of graphs, each graph describing one of the measures of document similarity, to compute a combined graph that describes the combined document similarity.

18. (Unchanged) A method as recited in Claim 1, wherein the combination of two or more measures of document similarity is achieved by taking the intersection of each of a plurality of graphs, each graph describing one of the measures of document

similarity, to compute a combined graph that describes the combined document similarity.

19. (Amended) A method as recited in Claim 1, further comprising the step of extracting ~~structural~~[17]**similarity**[18] information from the similarity matrix to obtain new documents supported by the set of training documents for each category.

20. (Amended) A method as recited in Claim 19, wherein the ~~structural~~[20]**similarity**[21] information is obtained by optimizing an objective function.

21. (Amended) A method as recited in Claim 19, wherein the ~~structural~~[24]**similarity**[25] information is obtained by only approximately optimizing an objective function.

22. (Unchanged) A method as recited in Claim 21, wherein approximately optimizing the objective function comprises repeated application of a growth transformation.

23. (Unchanged) A method as recited in Claim 19, further comprising the step of creating and storing a second matrix that represents an interim score for each document in each category.

24. (Unchanged) A method as recited in Claim 19, further comprising the steps of, periodically as the matrix is being computed, normalizing rows of the matrix by normalizing within each document, across all categories, whereby the score for one document in a particular category will depend on the scores for that document in all other categories.

25. (Unchanged) A method as recited in Claim 19, further comprising the steps of, periodically as the matrix is being computed, normalizing columns of the matrix by normalizing within each category, across all documents, whereby the score for one document in a particular category depends on the scores for all other documents in that category.

26. (Unchanged) A method as recited in Claim 1, in which the categories come from a manually defined taxonomy.

27. (Unchanged) A method as recited in Claim 1, wherein the categories are derived from logs of user queries.

28. (Amended) A method as recited in Claim 1, further comprising the steps of creating and storing ~~the~~[26] **a second**[27] matrix using columns representing documents and rows representing user sessions, and wherein values of elements of the second matrix represent interest in a document shown by a particular user in a particular session.

29. (Amended) A method as recited in Claim 1, further comprising the steps of creating and storing ~~the~~[28] **a**[29] matrix using columns representing user sessions and rows representing documents, and wherein values of elements of the second matrix represent interest in a document shown by a particular user in a particular session.

30. (Unchanged) A method as recited in Claim 28, wherein the element values are computed as a function of a time that a user has spent viewing a document associated with each element.

31. (Unchanged) A method as recited in Claim 28, further comprising the steps of creating and storing a second matrix representing a Similarity between pairs of documents i and j, wherein the second matrix is derived by comparing pairs of column vectors or row vectors, respectively i and j of the first matrix.

32. (Unchanged) A method as recited in Claim 28, further comprising the steps of creating and storing a second matrix representing a Similarity between pairs of documents i and j, by finding pairs of documents i and j which have high interest values for a particular user in a particular session or period of time.

33. (Unchanged) The method recited in Claim 1, further comprising the steps of:

identifying a category of a classification taxonomy of the hypertext system in which a first electronic document is presently classified; and

if a second electronic document is found to be highly Similar, storing information that classifies the second electronic document into the category.

34. (Amended) A computer-readable medium carrying one or more sequences of instructions, wherein execution of the one or more sequences of instructions by one or more processors causes the one or more processors to perform the steps of ~~categorizing a~~[30] **:**[31]

establishing[32] a plurality of training [33]**sets, wherein each training set is associated with a category and includes training documents that have been classified as belonging to said associated category;**[34]

determining how strongly each document of said[35] plurality of ~~new electronic~~ [36]documents ~~into a set of categories, each of~~[37]**corresponds to each of said plurality of categories by determining similarity between said each document and**[38] the ~~categories containing~~[39] ~~a plurality of training~~ [40]~~set documents, by~~[41]**documents that belong to the training set of said category; and**[42]

wherein the step of determining similarity is performed[43] using a matrix representing document similarity that is derived by combining two or more measures of document similarity.

35. Cancelled

36. Cancelled

37. Cancelled